Object Instance Retrieval in Assistive Robotics: Leveraging Fine-Tuned SimSiam with Multi-View Images Based on 3D Semantic Map

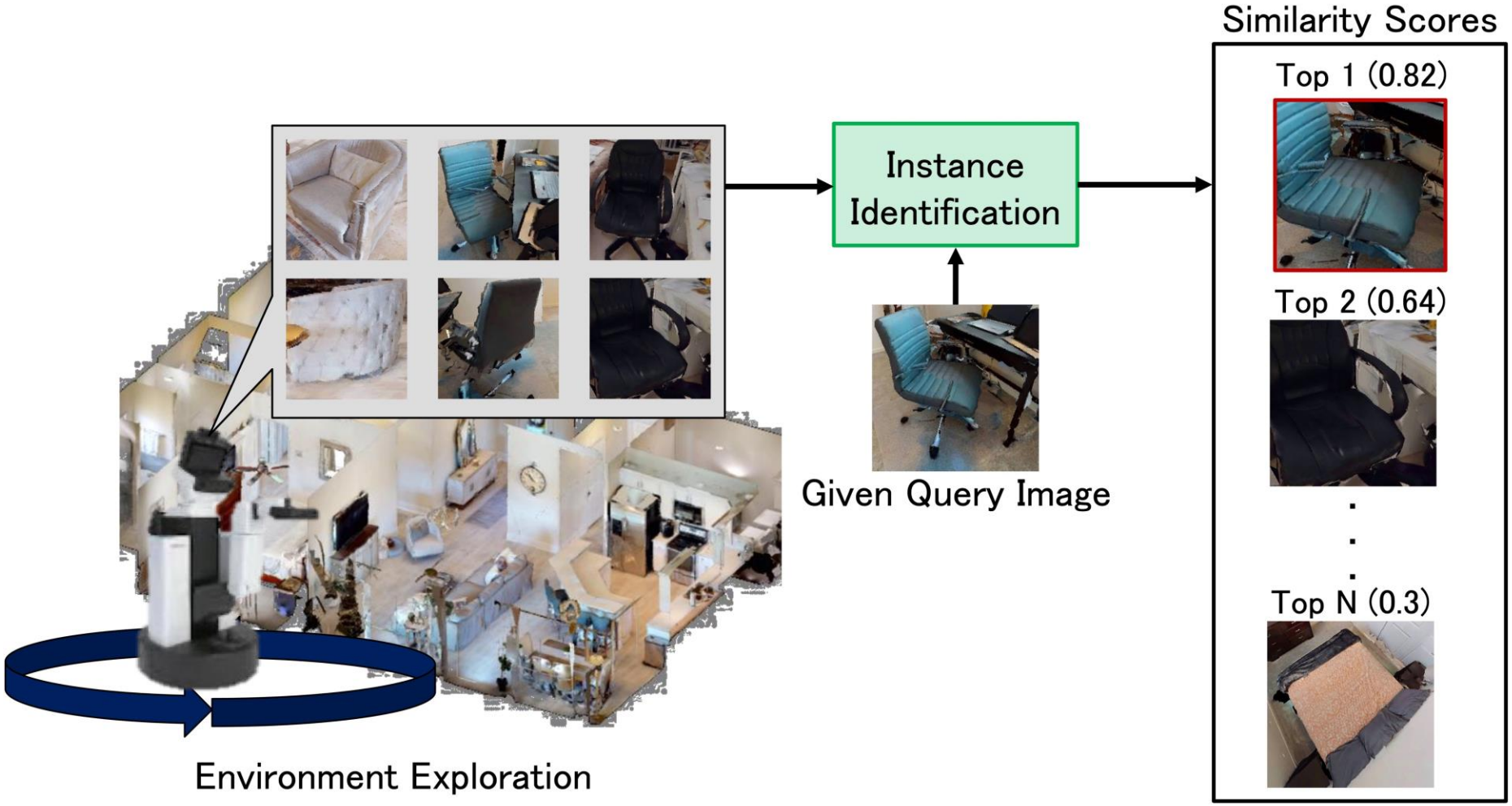
Taichi Sakguchi, Akira Taniguchi, Yoshinobu Hagiwara,
Lotfi El Hafi, Shoichi Hasegawa, Tadahiro Taniguchi

Ritsumeikan University, Japan

Task: Instance-specific Image Goal Navigation (InstanceImageNav)

InstanceImageNav:

The task of finding instances identical to a query image in an environment with multiple instances of the same object class, each with a different appearance.



Problem Statements:

Problem 1:

To solve InstanceImageNav, robots need to identify different instances of the same class of objects.

Contrastive language-image pre-training (CLIP), which was used in previous research on object search, is not suitable for fine-grained identification tasks.

Problem2 :

When a robot explores the environment and observes an object, it observes the object from various 3D viewpoints.

In this case, the appearance could be significantly different and could decrease the success rate of InstanceImageNav

Hypothesizes:

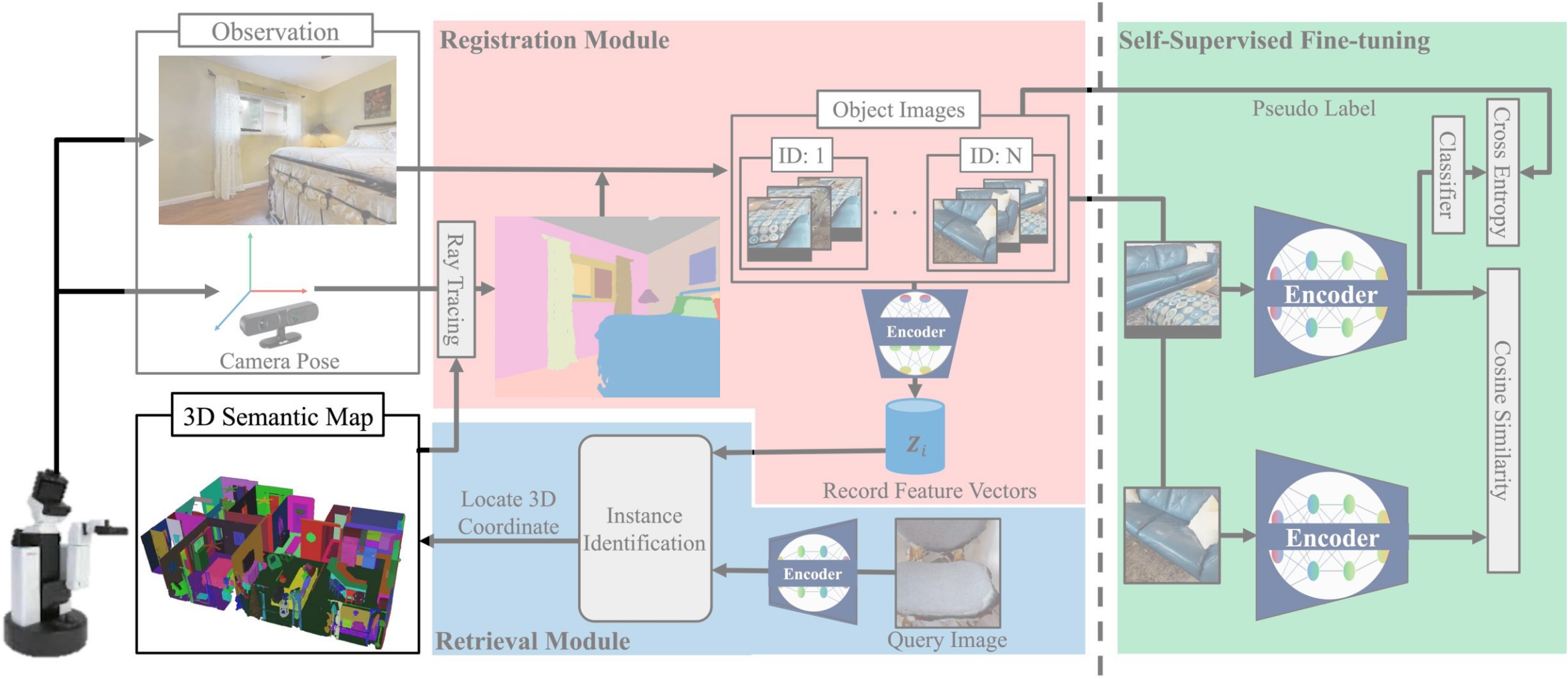
1. Models from contrastive learning between images could be superior to CLIP for instance-level object identification.

2. Increasing image similarity across various 3d viewpoints in self-supervised learning could improve instance retrieval accuracy.

Proposal: SimView

SimView:

Leveraging multi-view images based on a 3D semantic map of the environment and contrastive learning to train an instance identification model on-site



Simulation experiments: Quantitative result

Metrics: mean Average Precision

Comparison Method

- Unimodal Contrastive Learning Method
  - SimSiam, DINOv2, SimCLR
- Multimodal Contrastive Learning Method
  - CLIP

Method	Arch.	Env. 1	Env. 2	Env. 3	Env. 4	Env. 5	Env. 6	Env. 7	Env. 8	Env. 9	Avg.
SimView	ResNet50	**0.79**	**0.67**	**0.68**	**0.91**	0.68	**0.45**	**0.72**	**0.74**	**0.78**	**0.72**
SimSiam	ResNet50	0.70	0.58	0.56	0.80	0.65	0.41	0.63	0.68	0.75	0.64
DINOv2	ViT-B/14	0.51	0.48	0.45	0.66	**0.71**	0.41	0.55	0.44	0.61	0.54
SimCLR	ResNet50	0.65	0.56	0.52	0.83	0.64	**0.45**	0.66	0.62	0.75	0.63
CLIP	ResNet50	0.51	0.36	0.35	0.51	0.48	0.44	0.42	0.44	0.59	0.46
CLIP	ViT-B/16	0.56	0.38	0.37	0.47	0.48	0.37	0.38	0.35	0.48	0.43